View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Elsevier - Publisher Connector

JACC Vol. 14, No. 3
September 1989:44A–7A

Development and Analysis of Observational Data Bases

THOMAS H. LEE, MD, ScM, LEE GOLDMAN, MD, MPH
Boston, Massachusetts

Medical data bases have become an increasingly popular tool in clinical research. The observational data can be used to generate hypotheses for the development of clinical prediction rules and confirmation or extension of findings from other studies. When the analyses are carefully performed, the results from observational data bases generally have confirmed and complemented the findings of major randomized trials. In recent years, important clinical insights have been derived from analyses of data bases of large scale epidemiologic investigations; such insights may provide significant information on the impact of health policy decisions on patient care. To maximally and carefully use observational data, certain principles in both study design and the data collection and analysis phases should be applied.

(J Am Coll Cardiol 1989;14:44A–7A)

Role of Medical Data Bases in Clinical Research

Although randomized trials are accepted as the ideal methodology for evaluating a therapy, there are many clinical issues for which this experimental study design is inappropriate or impractical (1). For example, randomized trials are generally not useful for the identification of risk factors for adverse outcomes. When the goal of the investigation is the evaluation of an intervention, randomized trials may be impossible when the disease of interest is rare (for example, congenital heart diseases) or the outcome of interest occurs at a low rate because of statistical power limitations. Furthermore, even when a randomized trial has been performed, it may not resolve management issues for important subgroups of the study patients or other groups that may have been excluded from the study, such as the elderly, women or patients with a low ejection fraction.

Types of medical data bases. As a result, medical data bases have become an increasingly popular tool in clinical research (2). The observational data that are collected within these data bases have been used for exploratory (hypothesis-generating) analyses, the development of clinical prediction rules (3) and the confirmation or extension of findings from other studies (4). When analyses are carefully performed, the results generally have confirmed and complemented the findings of major randomized trials (4).

Several different types of data bases are currently being used in clinical investigations. The most common is the limited data base that is compiled for a single investigation, often by a single investigator. In contrast, many investigators are needed to start or maintain larger data bases such as the Duke University Cardiovascular Disease Databank (5) or the National Heart, Lung, and Blood Institute Percutaneous Transluminal Coronary Angioplasty Registry (6), which have been developed to capture the clinical experience with one set of diseases or a new technology over several years. These larger data bases can provide data for many investigations over a long period. Data that have been collected in the course of a randomized trial such as the Coronary Artery Surgery Study (CASS) Registry (7) can also be used for analyses that are only indirectly related to the original investigation.

In recent years, many investigators have taken advantage of data bases that have been compiled for purposes other than hospital-based research. For example, important clinical insights have been derived from analysis of the data bases of large-scale epidemiologic investigations such as the Framingham Heart Study (8) or the Nurses' Health Study (9). Similarly, health care researchers are using nonmedical data bases such as hospital fiscal or Medicare/Medicaid data bases, and analyzing these data separately or using them to complement independently collected clinical data. These studies may provide clinical findings or insights into the impact of health policy measures on patient care (10).

From the Division of Clinical Epidemiology, Consolidated Department of Medicine, Brigham and Women's and Beth Israel Hospitals, and the General Medicine and the Cardiovascular Divisions, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts. Dr. Lee is the recipient of a Public Health Service Clinical Investigator Award (HL-1594-01) from the National Heart, Lung, and Blood Institute, Bethesda, Maryland.

Address for reprints: Thomas H. Lee, MD, Division of Clinical Epidemiology, Consolidated Department of Medicine, Brigham and Women's and Beth Israel Hospitals, 75 Francis Street, Boston, Massachusetts 02115.

©1989 by the American College of Cardiology

0735-1097/89/$3.50

Principles of data base collection and analysis. Regardless of the type of data base, there are potential liabilities that must be considered when using observational data. These liabilities are magnified if the investigator was not involved in the data-gathering process because he or she may not be aware of which variables are unreliable. Furthermore, in the analysis of observational data, the relation between an exposure (for example, cardiac surgery) and an outcome (for example, death) is likely to be confounded by differences in the distribution of clinical characteristics between exposed and unexposed subjects.

To minimize these liabilities, investigators should apply certain principles in both the study design and data collection and analysis phases of an investigation. The underlying themes to these principles are: 1) ensuring that data are collected as objectively as possible, and 2) ensuring that comparisons are made between groups that are truly comparable. Some of these principles will be outlined in the following discussion.

Study Design and Data Collection

Prospective data collection. Studies that use observational data are often incorrectly perceived as retrospective by their very nature and, hence, susceptible to bias. In actuality, many observational data bases collect data prospectively; even data that are retrieved from a review of medical records may be recorded prospectively if the investigator uses an appropriate study design. Thus, if clinical information was entered into the medical record by a physician who did not know the patient's outcome, and if that information is recorded from the chart by a researcher who also did not know the patient's outcome status or the study hypothesis, the introduction of bias can be avoided.

The collection of prospective data on large numbers of patients can often be integrated into routine patient care. For example, in the Multicenter Chest Pain Study (11–14), we studied the ability of data from the emergency department evaluation of patients with acute chest pain to predict acute myocardial infarction and complications of ischemic heart disease that require intensive care unit management. In the collection of emergency department clinical data on approximately 15,000 patients at seven different hospitals, it was essential that the information be recorded without knowledge of the patient's actual outcome or diagnosis. If an investigator knows that a patient had had an acute myocardial infarction, he or she might be more likely to interpret an equivocal electrocardiogram as showing evidence of ischemia. The result of such a subtle bias might be an overestimation of the predictive value of the electrocardiogram.

Therefore, we sought the cooperation of the physicians in the emergency departments and the medical records departments at these hospitals, and arranged for a standardized study data form to be used as the official emergency department note. Instead of writing a text note on the blank sheet that usually serves as the emergency room form, the evaluating physicians entered clinical data on the study data form, which was placed in the medical record and substituted for the usual text. A copy of the form was also taken by the Chest Pain Study research assistants and used for data entry.

Patients were classified on the occurrence of outcome events such as a diagnosis of acute myocardial infarction on the basis of subsequent cardiac enzyme results and clinical events. In cases when the outcome was uncertain, other reviewers who did not know the emergency department data were asked to classify the patient.

With this study design, the clinical data could not be biased by knowledge of the patient's outcome, and the definition of outcomes could not be biased by knowledge of the clinical data.

This approach demonstrates the advantages of integrating data collection into routine care in large-scale studies. It would have been both expensive and impractical to have research assistants record emergency department clinical data from chart reviews or interviews, and it would probably have been impossible to ensure that the data were recorded without knowledge of patient outcome. The emergency department physicians benefited by having a convenient alternative to writing a full text note, and the Chest Pain Study was able to collect unbiased data at low costs. Although cost may not be an issue in investigations involving smaller numbers of patients, controlling potential bias is essential regardless of the size of the data base.

Objective, reproducible data recording. Even if data are collected prospectively, they may lose their value if they are not recorded in an objective and reproducible manner. Nonrandom error will occur if data are recorded differently for patients with various outcomes. For example, an investigator might be more likely to perceive ischemic changes on the electrocardiogram of patients who were known to have died, and thus distort the predictive value of such changes. In such situations, specific bias will be introduced.

Random error is introduced if a variable is recorded with a high degree of variability in all patient groups. In this situation, a true relation between that variable and the outcome might be overlooked; thus, the study would be biased toward the null hypothesis.

Clear definitions of variables and outcomes of interest. Investigators who have the opportunity to design their own data sets should, therefore, establish clear definitions of the variables and outcomes of interest. These definitions should be available in writing during data collection and analysis, and any changes or additions to the "code book" should be documented. A code book is essential if multiple reviewers perform the data collection, and useful even if only one person is involved.

With these definitions, multiple reviewers should be able to place patients in mutually exclusive categories reproduc-

JACC Vol. 14, No. 3
September 1989:44A–7A

ibly. These categories should reflect important clinical distinctions. For example, patients should not be classified simply as smokers or nonsmokers; multiple specific categories are needed to capture information on how much they smoke, how long they have smoked or how much time has passed since they have quit. Areas of potential uncertainty can often only be detected with pilot testing of a data form.

When investigators are working with data sets that have been collected by others, they should be sure that they know the precise definitions of variables. There is no substitute for collaboration with someone who was involved in the data collection because that person is most likely to know how questions were asked and which variables are unreliable.

Investigators who use data bases that have been gathered for purposes other than medical research, such as fiscal data bases, should also be aware that the accuracy of such data may not meet usual research standards. As an example, Iezzoni et al. (15) recently reported that 260 (26%) of 1,003 cases that had been coded as acute myocardial infarction by the medical records departments at 15 Boston hospitals did not meet clinical criteria for this diagnosis. Error rates in this range can be expected for other variables in nonresearch data bases. As just noted, if these errors are random, the analysis will be biased toward the null hypothesis; if they are nonrandom, a specific bias will be introduced.

Analysis

Avoiding adjusting for confounding factors. In analyzing observational data, investigators must always be concerned that an apparent association between a factor of interest (that is, the exposure) and outcome may result from confounding by a third factor that is associated with both the exposure and the outcome. For example, the observation that the mortality rate is higher among patients who are treated with digoxin after acute myocardial infarction than among patients who do not receive this medication (16) did not appear to be a function of deleterious effects of digoxin, but rather was reflective of greater severity of illness among the patients who received digoxin.

Several strategies may be employed to adjust for the potential impact of such confounding differences, but fundamental to all strategies is that the investigator cannot detect confounding if he or she does not look for it. Thus, investigators should collect and include in their analysis data on all variables that are potential confounders.

The simplest and most intuitive strategy for avoiding confounding is to match patients on the basis of potential confounding characteristics. Thus, to compare the rates of coronary artery disease in patients who do and do not drink alcohol, one could match patients in the two groups on the basis of factors that were risk factors for coronary heart disease, such as smoking status. This strategy would ensure that the role of smoking was the same in both the drinking and nondrinking groups and, therefore, remove its effect from the analysis.

The disadvantage of this approach is that the matching process becomes unwieldy when there are many potential confounders. In the preceding example, to control for confounding by other risk factors for coronary heart disease, patients would have to be matched not only for smoking status, but also for age, blood pressure, cholesterol levels, diabetes status, weight and family history.

Similarly, stratified analyses that divide patients into strata defined by potential confounders are simple and intuitive, but also suffer from the disadvantage of being impractical when there are many potential confounders. For example, if one creates a stratum for diabetic smokers, strata must also be created for diabetic nonsmokers, nondiabetic smokers and nondiabetic nonsmokers. Consideration of other confounders would lead to an exponential increase in the number of strata; hence, many strata would not have enough subjects in them for efficient or meaningful analysis.

Multivariate and parametric analysis. These limitations help explain the growing popularity of sophisticated multivariate techniques that are now widely available in software packages for microcomputers and minicomputers. For analyses with dichotomous end points (for example, dead or alive), parametric techniques such as multiple logistic regression (17,18) or linear discriminant analysis (19) can be used to identify the independent predictive importance of the exposure of interest and the potential confounders.

Although these techniques are powerful, they should be used only by investigators who are aware of the assumptions on which they are based. For example, with logistic regression analysis, interactions between covariates may not be detected if they are not anticipated in advance. Limitations of parametric techniques have led to interest in nonparametric multivariate techniques such as recursive partitioning (20). The relative merits of these approaches have been discussed elsewhere (21).

Another approach attempts to combine the features of cross-stratification and multivariate modeling by stratifying subjects according to their values from a model known variously as a "multivariate confounder score" (22) or a "propensity score" (23). A multivariate confounder score, as proposed by Miettinen (22), is derived by developing either an outcome model for the risk of the outcome as a function of the possible confounders and the exposure or, alternatively, an exposure model that describes the prevalence of the exposure in the study population as a function of the possible confounders and the outcome. In either case, the subjects are then stratified on the basis of their "score" according to the model; then, the association of the exposure and outcome is evaluated across these strata.

The propensity score. This score, as proposed by Rosenbaum and Rubin (23) for cohort studies, is developed from an exposure model and differs from the multivariate confounder

score by not containing a term for the outcome. Rosenbaum and Rubin (23) demonstrated that not including a term for the outcome is appropriate for cohort studies, and described how this approach might be preferable when more than one outcome is under consideration and when the outcome is ordinal or continuous in scale.

This approach has been used, for example, to clarify the relation between the use of digoxin and death after acute myocardial infarction (16). Nevertheless, the complexity of these combination methods has restricted their application in clinical investigation despite their intuitive appeal. Furthermore, the resulting strata, which are based on a "score" from a multivariate model, may possess little intrinsic meaning beyond representing different levels of risk or exposure prevalence. To address these problems, Cook and Goldman (24) recently described asymmetric stratification, a technique that uses recursive partitioning instead of a propensity score to develop strata to control for the relations between potential confounders and either the exposure or outcome of interest.

Conclusions. Regardless of the analytic strategy, one can never be sure that all potential confounders have been recognized. There may be unknown risk factors for an outcome that are distributed unevenly in the study population. As a result, analysis of observational data can never prove causation beyond a shadow of a doubt. However, reproducible, strong associations that demonstrate a dose-response effect may be sufficient to guide clinical practice or identify issues that are appropriate for randomized trials.

In this age of cost constraints on research of all types, the role of data base research in clinical investigation can be expected to expand. As this expansion occurs, methodologic standards for the development and analysis of observational data bases should become better defined and accepted.

References

1. Feinstein AR. An additional basic science for clinical medicine. II. The limitation of randomized trials. Ann Intern Med 1983;99:544–50.
2. Goldman L, Mushlin AI, Lee KL. Using medical databases for clinical research. J Gen Intern Med 1986;1:S25–S30.
3. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med 1985;313: 793–9.
4. Hlatky MA, Califf RM, Harrell FE Jr, Lee KL, Mark DM, Pryor DB. Comparison of predictions based upon observational data with the results of randomized controlled clinical trials of coronary artery bypass surgery. J Am Coll Cardiol 1988;11:237–45.
5. Rosati RA, McNeer JF, Starmer CF, Mittler BS, Morris JJ Jr, Wallace AG. A new information system for medical practice. Arch Intern Med 1975;135:1017–24.
6. Holmes DR, Holubkov R, Vlietstra RE, et al. Comparison of complications during percutaneous transluminal coronary angioplasty from 1977 to 1981 and from 1985 to 1986: the National Heart, Lung, and Blood Institute Percutaneous Transluminal Coronary Angioplasty Registry. J Am Coll Cardiol 1988;2:1149–55.
7. Principal Investigators of CASS and Their Associates. National Heart, Lung, and Blood Institute Coronary Artery Surgery Study. Circulation 1981;63(suppl I):I-1–81.
8. Kannel WB. Contributions of the Framingham Study to the conquest of coronary artery disease. Am J Cardiol 1988;62:1109–12.
9. Stampfer MJ, Willett WC, Colditz GA, Speizer FE, Hennekens CH. A retrospective study of past use of oral contraceptive agents and risk of cardiovascular diseases. N Engl J Med 1988;319:1313–7.
10. Fitzgerald JF, Moore PS, Dittus RS. The care of elderly patients with hip fracture. Changes since implementation of the prospective payment system. N Engl J Med 1988;319:1392–7.
11. Goldman L, Cook EF, Brand DA, et al. A computer protocol to aid in predicting myocardial infarction among emergency department patients with acute chest pain: prospective multi-center validation. N Engl J Med 1988;318:797–803.
12. Lee TH, Rouan G, Weisberg MC, et al. Patients with acute myocardial infarction sent home from the emergency room: clinical characteristics and natural history. Am J Cardiol 1987;60:219–24.
13. Lee TH, Rouan GH, Weisberg MC, et al. Sensitivity of routine clinical criteria for diagnosing myocardial infarction within 24 hours of hospitalization. Ann Intern Med 1987;106:181–6.
14. Beamer AD, Lee TH, Cook EF, et al. Diagnostic implications of the circadian variation of the onset of chest pain. Am J Cardiol 1987; 60:998–1002.
15. Iezzoni LI, Burnside S, Sickles L, Moskowitz MA, Sawitz E, Levine PA. Coding of acute myocardial infarction. Clinical and policy implications. Ann Intern Med 1988;109:745–51.
16. Muller JE, Turi ZG, Stone PH, et al. Digoxin therapy and mortality after myocardial infarction: experience in the MILIS Study. N Engl J Med 1986;314:265–71.
17. Cox DR. The Analysis of Binary Data. London: Chapman and Hall, 1970.
18. Breslow NE, Day NE. Statistical Methods in Cancer Research. The Analysis of Case-Control Studies, vol 1. Lyon: International Agency for Research on Cancer, 1980.
19. Kleinbaum DG, Kupper LL. Applied Regression Analysis and Other Multivariate Methods. North Scituate, MA: Duxbury Press, 1978.
20. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Belmont, CA: Wadsworth International Group, 1984.
21. Cook EF, Goldman L. Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. J Chronic Dis 1984;37:721–31.
22. Miettinen OS. Stratification by a multivariate confounder score. Am J Epidemiol 1976;104:609–20.
23. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effect. Biometrika 1983;70:41–55.
24. Cook EF, Goldman L. Asymmetric stratification: an outline for an efficient method for controlling confounding in cohort studies. Am J Epidemiol 1988;127:626–39.